The EMBO Journal vol.8 no.7 pp.1939 - 1946, 1989

Sequence analysis of α1(VI) and α2(VI) chains of human type VI collagen reveals internal triplication of globular domains similar to the A domains of von Willebrand factor and two α2(VI) chain variants that differ in the carboxy terminus

Mon-Li Chu, Te-cheng Pan, Dorothy Conway, Huey-Ju Kuo[1], Robert W.Glanville[1], Rupert Timpl[2], Karlheinz Mann[2] and Rainer Deutzmann[2]

Departments of Biochemistry and Molecular Biology and Dermatology, Jefferson Institute of Molecular Medicine, Thomas Jefferson University, Philadelphia, PA 19107, [1]Shriners Hospital for Crippled Children, Portland, OR 97201, USA and [2]Max-Planck-Institut für Biochemie, D-8033, Martinsried, FRG

Communicated by R.Timpl

Amino acid sequences of human collagen α1(VI) and α2(VI) chains were completed by cDNA sequencing and Edman degradation demonstrating that the mature polypeptides contain 1009 and 998 amino acid residues respectively. In addition, they contain small signal peptide sequences. Both chains show 31% identity in the N-terminal (~235 residues) and C-terminal (~430 residues) globular domains which are connected by a triple helical segment (335–336 residues). Internal alignment of the globular sequences indicates a repetitive 200-residue structure (15–23% identity) occurring three times (N1, C1, C2) in each chain. These repeating subdomains are connected to each other and to the triple helix by short (15–30 residues) cysteine-rich segments. The globular domains possess several N-glycosylation sites but no cell-binding RGD sequences, which are exclusively found in the triple helical segment. Sequencing of α2(VI) cDNA clones revealed two variant chains with a distinct C2 subdomain and 3′ non-coding region. The repetitive segments C1, C2 and, to a lesser extent, N1 show significant identity (15–18%) to the collagen-binding A domains of von Willebrand factor (vWF) and they are also similar to some integrin receptors, complement components and a cartilage matrix protein. Since the globular domains of collagen VI come into close contact with triple helical segments during the formation of tissue microfibrils it suggests that the globular domains bind to collagenous structures in a manner similar to the binding of vWF to collagen I.

Key words: alternative splicing/integrin/microfibrillar collagen/oligomer binding/polypeptide sequences/von Willebrand factor

Introduction

Collagen VI is a unique component within the family of collagenous proteins with an ubiquitous occurrence in soft connective tissues and cartilage. It forms a major class of tissue microfibrils that determines the principal function of the protein (Timpl and Engel, 1987; Burgeson, 1988). The special localization and functions are reflected in an unusual structure of collagen VI monomers, which consist of two large globular domains connected by a short (105 nm) triple helical segment. Constituent polypeptides chains are α1(VI) and α2(VI) of ~140 kd and α3(VI) of ~250 kd (Trüeb and Winterhalter, 1986; Colombatti *et al.*, 1987) which occur in stoichiometric proportions (Jander *et al.*, 1983; Chu *et al.*, 1987). The three chains of human collagen VI were recently cloned (Chu *et al.*, 1987), assigned to chromosomes 2 and 21 (Weil *et al.*, 1988) and shown to be differentially regulated in cell culture and under the influence of cytokines (Hatamochi *et al.*, 1989; Heckmann *et al.*, 1989).

Immunoelectron microscopy has indicated that collagen VI forms a fibrillar network independent from the major fiber-forming collagens (von der Mark *et al.*, 1984; Keene *et al.*, 1988). However, frequent contacts are observed between both fibrous systems as well as with basement membranes (Bruns, 1984; Bruns *et al.*, 1986; Keene *et al.*, 1988). The microfibrils were also shown to be close to cells, supporting other observations that collagen VI acts as a cell-binding protein *in vitro* (Carter, 1982). Major cell-binding activities could be localized to triple helical segments of the α2(VI) and α3(VI) chains and blocked by synthetic RGD-containing peptides (Aumailley *et al.*, 1989). It was also shown that integrin receptors are involved in this recognition (Wayner and Carter, 1987). The major functions of collagen VI seem, therefore, to be the anchorage of cells to the extracellular matrix and of some basement membranes to their underlying stroma. The assembly of such supporting fibrils is a rather complex process and includes defined oligomer intermediates (Furthmayr *et al.*, 1983; Engvall *et al.*, 1986).

The potential biological roles of collagen VI are still insufficiently correlated to particular structural features. The triple helical sequence of all three chains of human collagen VI was recently determined and showed the presence of functional cysteines and RGD sequences (Chu *et al.*, 1988). A partial sequence of chick α2(VI) chain, including portions of the globular domains, was also reported (Trüeb *et al.*, 1989). Here, we have determined the whole sequence of human α1(VI) and α2(VI) chains showing repetitive units within the globular domains. These units are similar to domains in von Willebrand factor (vWF) for which collagen binding has been shown (Titani and Walsh, 1988). The findings introduce new considerations about the binding potential and function of collagen VI.

Results

Sequence of the N- and C-terminal globular domains of α1(VI) and α2(VI) chains deduced from cDNA

We have previously isolated several cDNA clones from human fibroblasts (F) and placenta (P) libraries and have shown by partial sequencing and restriction enzyme mapping that they encode various overlapping portions of the α1(VI) and α2(VI) chain (Chu *et al.*, 1988). In addition, two clones

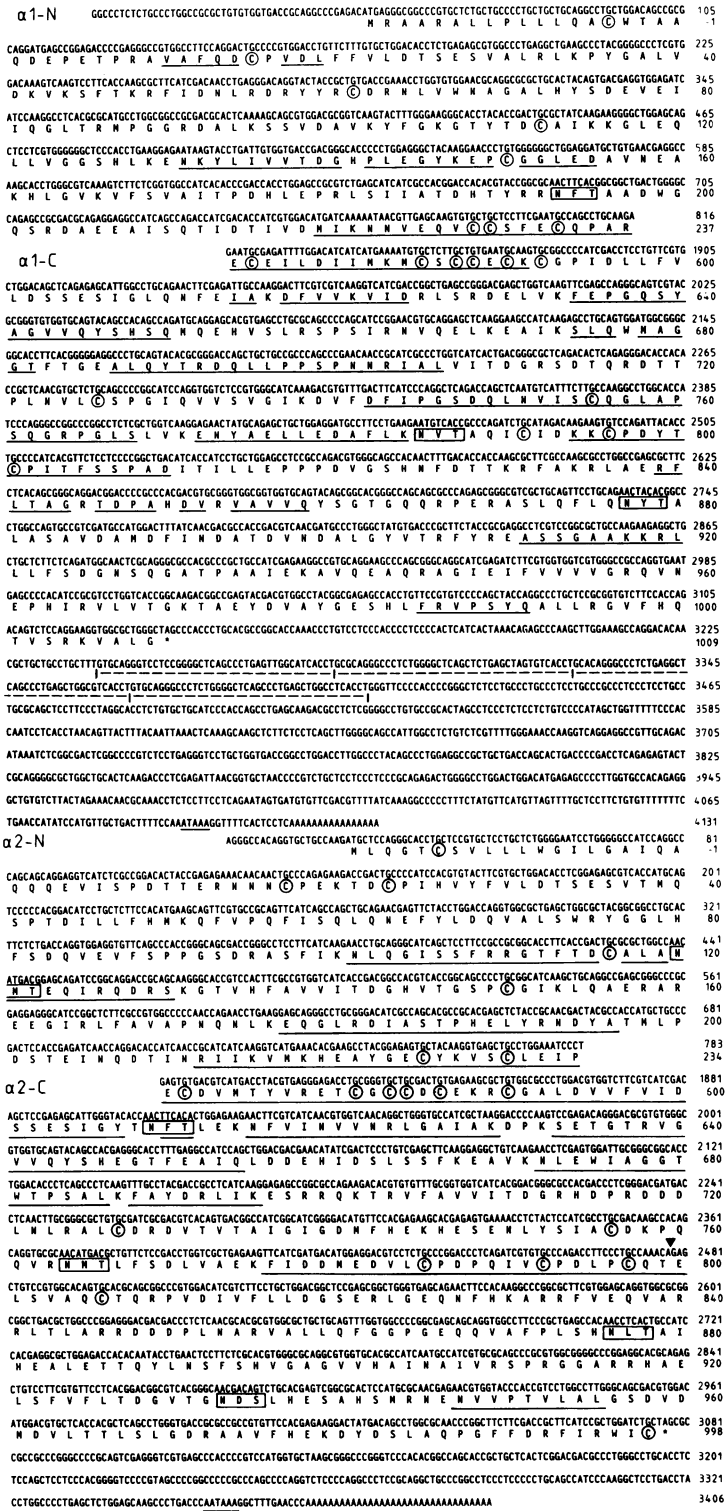

Fig. 1. Nucleotide and deduced amino acid sequences of the N- and C-terminal globular domains of human collagen α1(VI) and α2(VI) chains. Numbering of the sequence is continuous including 1005–1008 nt encoding the triple helical sequence located between both globular domains (Chu *et al.*, 1988). A possible alternative splice site in α2(VI)-C is denoted by a black arrowhead. Potential polyadenylation signals in the 3′ non-coding regions are underlined. Cysteine residues are encircled and potential N-linked carbohydrate acceptor sites are shown in boxes. Sequences confirmed by Edman degradation of peptides are underlined. Differences between deduced amino acid sequence and results of Edman degradation (given in brackets) were noted in α1(VI) chain for position 16(R) and in α2(VI) chain for positions 586(L), 599(F), 608(L) and 946(Q). They are in all cases explained by single base substitutions. A 4-fold repeat of 42 nt in the 3′ non-coding region of α1(VI)-C is marked by broken lines.

corresponding almost to the full mRNA length were obtained for α1(VI) (F157, 4.1 kb) and α2(VI) chain (F225, 3.4 kb). Sequence analysis of these clones demonstrated open reading frames of 3084 nt for α1(VI) and of 3054 nt for α2(VI) mRNA, non-coding regions of variable length at the 3′ end and small non-coding segments at the 5′ end (Figure 1). The 3′ ends terminated in a poly(A) tail closely preceded by typical polyadenylation signals. The predicted initiation

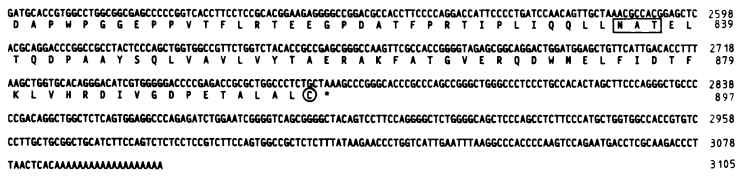

Fig. 2. Nucleotide and deduced amino acid sequence of a variant 3′ segment of α2(VI) chain. This sequence starts at nt 2479 of the α2(VI) sequence (Figure 1) and follows its numbering.

codons were also preceded by a typical consensus sequence (Kozak, 1987) including a purine base 3 nt upstream. The 3′ non-coding region of α1(VI) but not of α2(VI) showed shortly after the stop codon an interesting 4-fold, highly identical repeat of 42 nt (Figure 1). Similar repeats have so far not been found in other DNA sequences.

The cDNA sequence allowed us to deduce the entire amino acid sequence of the N-terminal and C-terminal globular domains (Figure 1) and of the triple helical segments (Chu *et al.*, 1988; see Figure 4) of both collagen VI chains. The sequences start with a characteristic signal peptide (von Heijne, 1985). Predictions of their cleavage sites (von Heijne, 1986) indicate that the signal peptides possess 19 (α1) and 20 (α2) residues respectively, and that the mature polypeptides contain 1009 (α1) and 998 (α2) residues. They start with a QDE (α1) or QQQ (α2) sequence. The sizes of both N-terminal globular domains are similar (237 versus 234 resiudes) as is the case for the 2-fold larger C-terminal globular domains (436 versus 429 residues). Other characteristic features of the globular domains are the presence of 6−7 N-terminal and 11−13 C-terminal cysteines in each chain and of eight potential N-glycosylation sites (NXT/S) (Figure 1).

A variant α2(VI) chain that differs in the carboxy terminus

Restriction enzyme mapping of various α2(VI) cDNA clones showed a striking variation in the number of *Pst*I recognition sites. Based on the digestion pattern, the cDNAs can be divided into one group of clones (F225, F126, P30, P102) containing two *Pst*I sites at their 3′ ends (Chu *et al.*, 1988), whereas the other group of clones (P1, P16, P201) lacked these two sites. Clone P1 (2.5 kb) from the second group was completely sequenced and the nucleotide sequence was identical to that of F225 at the 5′ end from nt 730 (start of P1) up to nt 2478 within the C-terminal globule; but thereafter, the sequence was entirely different (Figure 2). The variant segment C2a is 301 nt shorter than the corresponding α2(VI) sequence (Figure 1) with an in-frame deduced amino acid sequence of 98 residues. This reduces the size of the C-terminal globule to 328 residues in the predicted α2(VI) chain variant. An oligonucleotide primer close to the constant 3′ end sequence was used to sequence the variable region of all other clones and showed, consistent with the restriction mapping data, the typical α2(VI) sequence for F126, P30 and P102 and the variant sequence for P16 and P201.

Collagenase-resistant fragments and peptide sequences of α1(VI) and α2(VI) chain

Cleavage of a mixture of α1(VI) and α2(VI) chains with bacterial collagenase has been previously shown to generate three segments with a size of ~30−70 kd (Jander *et al.*, 1984; Gibson and Cleary, 1985; Trüeb and Winterhalter, 1986). They very likely correspond to the globular domains of the chains described here and, therefore, a similar digest of the chains was prepared from human amnion. It contained three major (GRBC1, 2 and 4) and one minor (GRBC3) peptide band (28−62 kd) when examined by electrophoresis. These fragments were purified to apparent homogeneity (Figure 3) and subjected to Edman degradation. GRBC4 failed to release any N-terminal amino acids, which was also the case when a mixture of intact α1(VI) and α2(VI) chains was examined. Sequence analysis of small peptides generated from GRBC4 (see below), however, clearly showed that it contained the N-terminal globular domains of both chains. GRBC1 and GRBC2 were identified in a similar way as the C-terminal globular domains of α2(VI) and α1(VI) chains respectively (Table I). The sizes of the identified GRBC fragments corresponded well with that predicted from cDNA sequence analysis (Table I), indicating that they comprise whole globular segments. The minor fragment GRBC3 could not be identified and may represent an α3(VI) chain fragment.

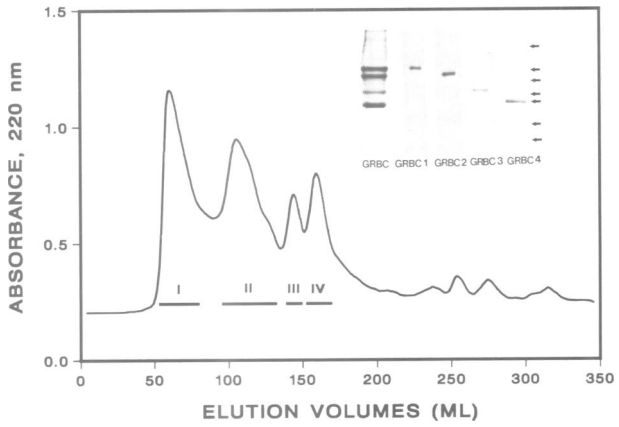

Fig. 3. Separation of large collagenase-resistant peptides of α1(VI) and α2(VI) chains by molecular sieve HPLC on two Bio-Sil TSK-250 and two Bio-Sil TSK-125 columns connected in tandem. Horizontal bars indicate pools used in further studies. Peak I contains trace amount of a large peptide generated from α3(VI). Peak II contained peptides GRBC1 and 2 which were further separated on a poly F reverse-phase column (not shown). Peak III and peak IV contained GBRC3 and 4 respectively. The insert is the electrophorogram of SDS−polyacrylamide gels (10−12.5%), which shows the starting material (GRBC) and the purified peptides GBRC1, 2, 3 and 4. The arrows indicate the globular protein standards phoshorylase (95 500), glutamate dehydrogenase (55 000), ovalbumin (43 000), lactate dehydrogenase (36 000), carbonic anhydrase (29 000), lactoglobulin (18 400) and cytochrome C (12 400).

Proteolytic cleavage of GRBC fragments followed by HPLC separation allowed the sequence analysis of 26 peptides (~370 positions), which could all be located to

various portions of the four globular domains (Figure 1). In addition, three peptides were obtained which belonged to the N-terminal globule of α3(VI) chain. This demonstrated that the mixture of α1(VI) and α2(VI) chains, which was prepared as in other studies (Timpl and Engel, 1987), contained a fragment of α3(VI) chain.

Comparison of amino acids sequences

Alignment of the entire α1(VI) and α2(VI) chain sequences revealed a striking homology in the peptide chains (Figure 4). Identical residues were found in 31% of the sequence of the N-terminal and C-terminal globular domains and in 37% of the triple helical segments, together with 15–20% conservative replacements. However, if the comparison of the triple helix is restricted to those positions not occupied by the glycine residues, then the identity drops to ~12%.

Internal comparison of the globular sequences within each chain revealed the existence of a 190- to 200-residue repeat as illustrated for the α1(VI) chain in Figure 5. This indicates the presence of three subdomains N1, C1 and C2, which may have arisen by triplication from a primordial segment. About 19–22% identical residues were found in the comparison of C1 and C2 from both chains. A comparable identity (19–23%) was found between N1 and C1 and was slightly lower (15–19%) between N1 and C2. The alignment scores were in the range 4–7 SD, demonstrating that this identity has not occurred by chance. Each of these subdomains is either preceded or terminated by short segments (15–30 residues) that have little in common except a variable number (3–6) of cysteine residues. The shorter sequence (C2a, 98 residues) of the α2(VI) chain (Figure 2) lacked the typical C2 repeat. C2 and C2a share only 10% identical residues, but the presence of ~30% conservative replacements indicates that they are remotely related.

The comparison of the globular domains with other protein sequences shows a particularly striking similarity to the collagen-binding A domains (Titani *et al.*, 1986) of vWF (Figure 5). These domains are also repeated three times in vWF (A1, A2, A3; Titani and Walsh, 1988) and are ~200 amino acid residues long. The best homology was found for the A1 domain showing 14–17% identical residues with N1, C1 and C2 of α1(VI) chain equally distributed over the entire length of the segments. The identity scores are similar for C1 and C2 (17%) of α2(VI) chain and insignificant (8%) for its N1 segment. With α2(VI)-C1 and -C2 there is a clear change in similarity over the entire length: ~45% identity in the first 40 residues, dropping to ~9% in the last 100-residue segment. As discussed elsewhere (Pytela, 1988)

Table I. Properties of collagenase-resistant peptides (GRBC) of human collagen α1(VI) and α2(VI) chains and their origin from N and C terminus

Peptide	Origin	Molecular mass (kd)[a]	
		Found	Calculated
GRBC1	α2-C	62	57
GRBC2	α1-C	51	53
GRBC3	–	36	–
GRBC4	α1-N + α2-N	28	28

[a]Determined by electrophoresis with globular proteins as markers and calculated from the sequence adding 2.5 kd for each N-glycosylation site.

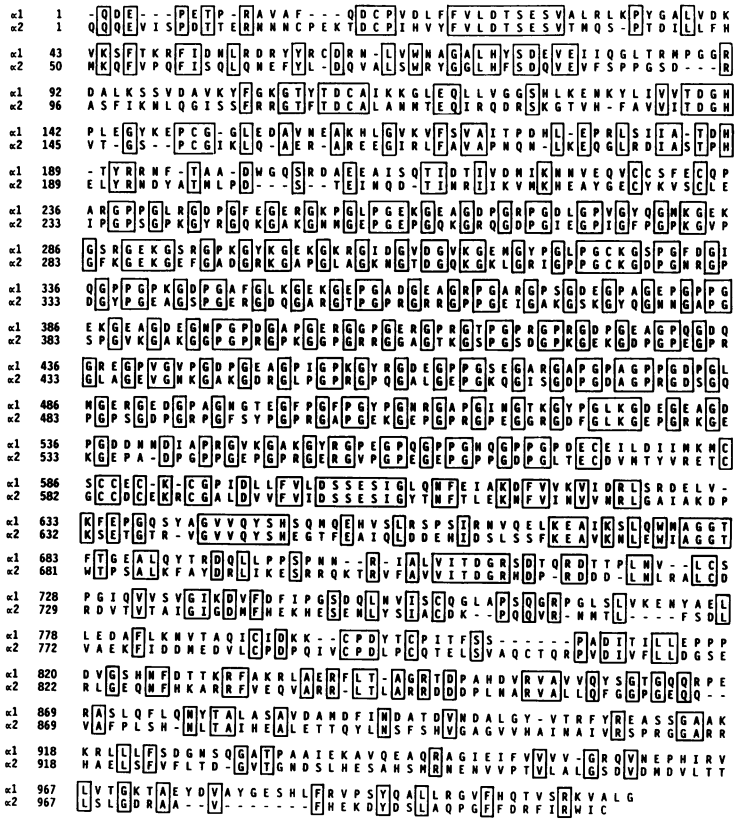

Fig. 4. Comparison of the complete amino acid sequences of α1(VI) and α2(VI) chain excluding their signal peptides. A few gaps were introduced in each sequence to maximize homology. Identical residues are boxed. The alignment score is 51.1 SD. Compiled from data in Figure 1 and Chu *et al.* (1988) for the triple helical segments.

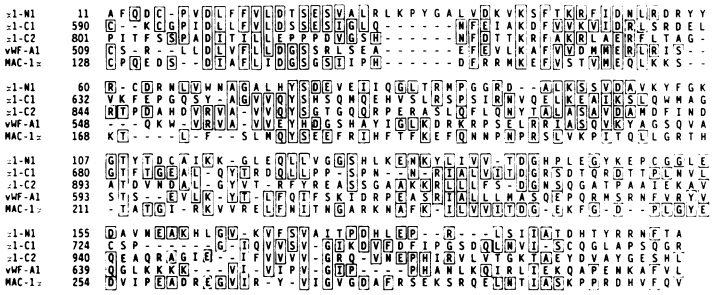

Fig. 5. Sequence comparison of the N1, C1 and C2 subdomains of α1(VI) chain with the A1 domain of von Willebrand factor (Titani *et al.*, 1986) and a segment of Mac-1 α chain (Pytela, 1988). Residues identical in at least two sequences are boxed. Alignment scores for the comparison C1−C2 were 5.69 SD, for C1−N1 6.81 SD, for C2−N1 4.28 SD, for C1−A1 5.43 SD, for C2−A1 5.59 SD and for N1−A1 5.02 SD.

the vWF A1 domain exhibits significant identity (~17%) with an α-chain-inserted domain of some integrin receptors (Mac-1, p150,95), complement components B and C2 and with a cartilage matrix protein (Argraves *et al.*, 1987). The C1 and C2 and to a lower degree the N1 segments of both collagen VI chains show a similar homology with these proteins. Two amphiphilic sequences, which in α2(VI)-C1 were VVFVIDSSESIG and FAVVITDGR, seem to be particularly conserved in all these proteins.

The N1 segment of α1(VI) but not of α2(VI) chain contains a 40-residue stretch that has 30% identical residues with ICAM-1, a cell-recognition protein belonging to the immunoglobulin gene super-family (Staunton *et al.*, 1988). Another similarity (20% identical residues) exists between N1 of the α2(VI) chain and the β-chain of S-100. This is a small brain protein that binds calcium and cytoskeletal proteins (Kligman and Hilt, 1988) and the gene is located close to those of α1(VI) and α2(VI) chains on chromosome 21 (Allore *et al.*, 1988). For both cases, the segments involved are located within the C-terminal 100 residues of N1, in a region that does not show good homology to vWF.

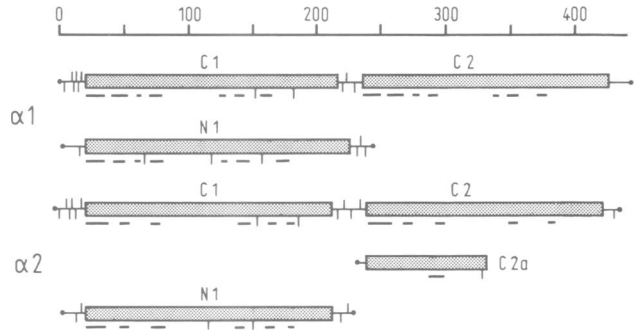

Fig. 6. Subdomain models of the N-terminal (N1) and C-terminal (C1, C2) globules of α1(VI) and α2(VI) chains. Horizontal stippled bars denote internally repeated segments that are also similar to vWF A domain with thick lines underneath outlining positions of maximal homology. They are connected by small variable segments (thin lines). C2a refers to an alternatively spliced segment. Positions of cysteines are indicated by short vertical lines. The scale on top gives the amino acids numbers. The C-termini of all structures are to the right-hand site.

Discussion

The cDNA sequences reported here complete the primary structure of collagen VI α1(VI) and α2(VI) chains, demonstrating that the mature polypeptides contain 1009 and 998 amino acid residues respectively. The sizes agree with a molecular mass of 110−140 kd determined for these chains in several previous studies by electrophoresis (Timpl and Engel, 1987). The entire nucleotide sequences determined including the non-coding regions were close to 4.2 kb (α1) and 3.5 kb (α2), which are the sizes previously determined for the respective mRNAs (Chu *et al.*, 1987). About 30% of the deduced amino acid sequence was supported by Edman degradation of peptides with only a few minor disagreements. The data also show that ~25% of the total mass of the chains is contributed by an N-terminal and ~40% by a C-terminal globular domain in agreement with the dumb-bell model proposed for collagen VI monomers (Furthmayr *et al.*, 1983; Engel *et al.*, 1985). Both globular domains are connected by a triple helical segment of 335 (α2) or 336 (α1) residues (Chu *et al.*, 1988).

The α1(VI) and α2(VI) chains share ~33% identical residues when both sequences are aligned with a small number of gaps. In addition, an internal repeat of 190−200 residues with ~20% identity exists once in the N-terminal and twice in the C-terminal globules of both chains. This indicates that they are in fact organized in three subdomains N1, C1 and C2 with a rather invariant pattern of small sequence segments of high similarity (Figure 6). Together, the data suggest that the internal repeats became triplicated on a single primordial gene before this gene became duplicated to give rise to the α1(VI) and α2(VI) chains. In the human genome both genes are in close proximity on chromosome 21 (Weil *et al.*, 1988) and presumably not separated by >200 kb (Cutting *et al.*, 1988).

Each chain contains 19−20 cysteine residues with a single one contributed by the triple helical domain. The remaining cysteines are distributed in some irregular fashion among the globular segments, with a few clusters found in short segments connecting the triple helix and the internal repeats (Figure 6). This indicates that not all are involved in stabilizing the globular folding by disulfide bridges. Various studies have in fact shown that the cysteines serve several purposes. Those in the clusters at the C-terminal end of N1 and the N-terminal end of C1 are likely to form in part intramolecular bridges between the chains which protect the triple helical domain from proteolytic attack (Jander *et al.*, 1983; Odermatt *et al.*, 1983). Two more cysteines are required to form disulfide bridges between the C-terminal globule and the triple helix in the staggered collagen VI dimers (Furthmayr *et al.*, 1983; Chu *et al.*, 1988). It is also known

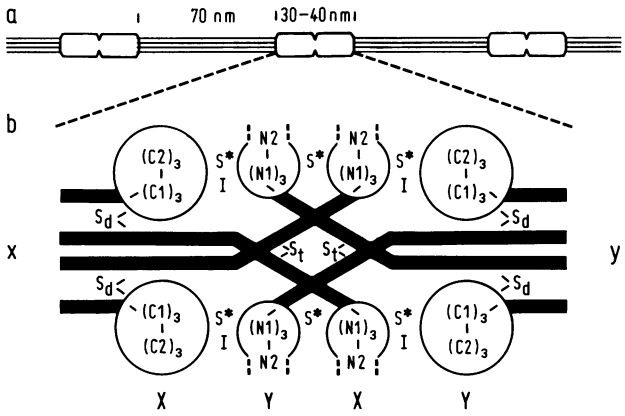

Fig. 7. Schematic model of collagen VI microfibrils (**a**) and enlarged schema of the junctional complex formed from two (x,y) overlapping collagen VI tetramers (**b**). Globular domains are denoted by N and C and refer to the structures shown in Figure 6 including N1, C1 and C2 from the α3(VI) chain. N2 indicates a large extra domain contributed exclusively by α3(VI) chain. The thick black lines represent the triple helical domains. The positions of disulfide bonds linking collagen VI dimers (S_d) and tetramers (S_t) are indicated (Furthmayr *et al.*, 1983; Chu *et al.*, 1988). Further disulfide bonds (S*) are known to link globular domains but their positions are tentative. I indicates sites of possible non-covalent interactions between the globules. The schema is modified from Furthmayr *et al.* (1983), Bruns (1984) and Engel *et al.* (1985) according to sequence data.

that collagen VI tetramers form tissue microfibrils by an overlapping end-to-end association (Figure 7), and that these may become stabilized by an unknown number of disulfide bridges. It is likely that they are contributed by both the N- and C-terminal globular domains. Because of this complexity the localization of disulfide bonds in collagen VI will remain a considerable but important task.

The presence of eight potential N-glycosylation sites in the globular domains suggests, as for the triple helical domain (Chu *et al.*, 1988), modification by branched oligosaccharides. Some of the sites are certainly utilized since we failed to identify, by Edman degradation, the accepting asparagine residue in two positions (120, 609) of α2(VI) chain. The globular domains lack any RGD sequence known to contribute putative cell-binding sites (Ruoslahti and Pierschbacher, 1987). However, the triple helical domain contains 11 of these RGD sequences (Chu *et al.*, 1988). They are functional in the α2(VI) and α3(VI) chains, as has been demonstrated by cell adhesion and inhibition assays (Aumailley *et al.*, 1989). This indicates that the triple helix and the globules of collagen VI are involved in different biological activities.

Studies with several α2(VI) cDNA clones provided convincing evidence for two variants, which could arise by alternative splicing at the 3′ end of α2(VI) mRNA involving portions of the coding and the entire non-coding region. Alternative splicing is supported by Southern blot analyses of genomic DNA, which suggested that the α2(VI) gene is present as a single copy in the human genome (Cutting *et al.*, 1988; and our unpublished results). Alternative splicing has been observed for a variety of different proteins (for a review see Breitbart *et al.*, 1987) including collagen IX, which is present in chick cornea and cartilage with different N-terminal domains (Svoboda *et al.*, 1988), and collagen XIII, which has several splice sites in the region encoding the triple helix (Tikka *et al.*, 1988). Whether the α2(VI) chain alternatives are tissue specific is unknown, but cDNA clones comprising more than one variant have been obtained from skin fibroblasts and placenta. The splicing eliminates entirely the C2 subdomain and a short segment containing one cysteine: these are replaced by a 101-residue shorter C2a subdomain with only little homology to C2. The functional and structural consequences of this variation remain to be studied. This also indicates a second α2(VI) mRNA species of ~3.2 kb. Such a smaller species has been detected as a minor band in Northern hybridization of skin fibroblast mRNA (Chu *et al.*, 1987).

Collagen VI also contains an α3(VI) chain in stoichiometric amounts (Jander *et al.*, 1983; Trüeb and Winterhalter, 1986; Chu *et al.*, 1987) that is twice the size of the α1(VI)/α2(VI) chains (Colombatti *et al.*, 1987) and provides essential triple helical structures for the formation of twisted dimers and tetramers (Chu *et al.*, 1988). Down-regulation of α3(VI) mRNA by γ-interferon has also been shown to be rate-limiting for the assembly, secretion and matrix deposition of collagen VI (Heckmann *et al.*, 1989). We have cloned a substantial portion of α3(VI) mRNA (Chu *et al.*, 1988) and shown by sequence anlaysis that it encodes similar domains N1, C1 and C2 as found in the other collagen VI chains (M.-L.Chu *et al.*, unpublished observations). In addition, the N-terminal globule of α3(VI) contains extra domains that may have a mass of 140 kd as judged from the size of a collagenase-resistant peptide (Trüeb and Winterhalter, 1986). Thus, the α3(VI) chain not only contributes unique structural elements to the triple helix but also to the globular domains. The α3(VI) chain could also exist as spliced variants since the mRNA appears as 2–4 distinct bands of ~8.5 kb (Chu *et al.*, 1987).

Globular domains are not a unique feature of collagen VI. They are found in the precursor forms of the fiber-forming collagens I, II and III (Kühn, 1984) and in the C-terminal domain NC1 of basement membrane collagen IV (Weber *et al.*, 1984). We could not detect any sequence similarities between these structures and the collagen VI globules. Yet, in both the α1(IV) and α2(IV) chains the NC1 domain consists of a repeating 100-residue segment with a high identity score (50–70%) including all cysteines in invariant positions (Oberbäumer *et al.*, 1985; Schwarz-Magdolen *et al.*, 1986). As in collagen IV, the globular domains of collagen VI do not represent precursor structures as shown by peptide sequences found at the extreme N and C termini (Figure 1) in tissue-extracted α1(VI)/α2(VI) chains. This demonstrates no, or only minimal, processing of both collagen VI chains except for the removal of signal peptides, which was predicted from pulse-chase experiments in cell cultures (Colombatti and Bonaldo, 1987). The α1(VI) and α2(VI) chains have blocked N termini like the procollagen I and III chains, where it is due to cyclic glutamine residues (Rohde *et al.*, 1979; Brandt *et al.*, 1984). An N-terminal glutamine is in fact predicted for both the α1(VI) and the α2(VI) chains (Figure 1).

The most surprising observation was the distinct homology between collagen VI subdomains C1 and C2, and to a lesser extent, of N1 to the A domains of vWF. Domains with a comparable identity (17–23%) to the A domain are found in some complement components (B, C2), a cartilage protein (Agraves *et al.*, 1987) and in the α-chains of integrins Mac-1, p150 and LFA-1 (Pytela, 1988; Corbi *et al.*, 1987) which belong to the β2 subfamily of these cellular receptors (Hynes, 1987). These integrins are mainly involved in

cell–complement and cell–cell interactions, with LFA-1 binding to a distinct cell-surface protein, ICAM-1 (Staunton *et al.*, 1988). Most other integrins, however, lack the A domain analog and are mediating cell–matrix interactions (Hynes, 1987; Ruoslahti and Pierschbacher, 1987). The A domains in vWF and complement B/C2 are connected to a cell-binding RGD sequence present in other domains which was considered to indicate a bridging function between cells and extracellular components for these proteins (Pytela, 1988). Collagen VI apparently belongs to this group of proteins since it shows RGD-dependent cell adhesion (Aumailley *et al.*, 1989). The ligands binding to A domains have so far only been identified for vWF and include fibrillar collagen I, heparin and the platelet surface protein Ib (Fujimura *et al.*, 1986, 1987; Pareti *et al.*, 1986; Titani and Walsh, 1988).

The presence of A-like domains in the α1(VI) and α2(VI) chains strongly suggests that they have a binding function but does not necessarily identify the nature of their ligands. Yet, it is conceivable that the triple helical domain of collagen VI itself could be mainly involved. The C-terminal globular domain contacts a particular region of the triple helix during dimer formation but selective reduction of a few disulfide bonds results in dissociation of this structure (Furthmayr *et al.*, 1983; Odermatt *et al.*, 1983; Jander *et al.*, 1984; Chu *et al.*, 1988). This indicates that the non-covalent interaction is of low affinity and needs disulfide bridges for stabilization. Another contact possibility exists in the junctional complex of microfibrils where the overlap between collagen VI tetramers produces a tight cluster of eight globular domains and triple helical segments (Figure 7). Here, even weak interactions may become potentiated by co-operative binding. It is in this context that, as shown by preliminary data (Rand *et al.*, 1988), collagen VI itself may also be a ligand for vWF binding. Interactions could of course also occur between globular domains to which N1 structures with similarity to ICAM-1 and S-100 may contribute. The reaction potential could in addition include heterotypic binding to collagens I and/or IV since collagen VI microfibrils are occasionally seen in contact with morphological structures containing these components (von der Mark *et al.*, 1984; Bruns *et al.*, 1986; Keene *et al.*, 1988). The molecular elucidation of this interaction repertoire is a challenge for future studies and may be facilitated by the sequence analysis shown here.

Materials and methods

Isolation of cDNA clones and nucleotide sequencing

Several cDNA clones encoding α1(VI) and α2(VI) chains have previously been isolated from a λgt11 library of human placenta and a λZAP library of human skin fibroblats (Chu *et al.*, 1988). Restriction fragments from selected clones (F157, F113, F225, P1, P6, P18) were subcloned into M13mp8 and mp19 vectors. The nucleotide sequences were obtained by dideoxy-chain termination (Sanger *et al.*, 1977) using sequenase sequencing kit (US Biochemicals, Cleveland, OH) with either M13 universal primer (Amersham, Arlington Heights, IL) or specific oligonucleotide primers derived from internal cDNA sequences. Some sequences were determined using fluorescent M13 primers (Applied Biosystems, Foster City, CA) and sequenase (US Biochemicals) according to protocols suggested by the suppliers. The samples were analyzed in an automated DNA sequencer (Model 370A, Applied Biosystems).

Purification of collagenase-resistant peptides

Intact type VI collagen was purified from human amnion (Kuo *et al.*, 1989) and dissolved in 0.5 M Tris–HCl, pH 7.5, 0.2 M NaCl, 8 M urea, 5 mM EDTA. It was reduced with 0.09 M 2-mercaptoethanol (16 h, 20°C) followed by alkylation with 0.09 M 4-vinylpyridine. The reduced sample was chromatographed on two Bio-Sil TSK-400 columns (21.5 × 600 mm), which were connected in tandem and equilibrated with 0.04 M Tris–acetate, pH 6.8, 6 M urea, 0.1 M sodium sulfate (TSK buffer). The peak containing the 140 kd α1(VI) and α2(VI) chains was dialyzed agianst 0.5 M Tris–HCl, pH 7.4, 0.4 M NaCl, 10 mM $CaCl_2$ containing protease inhibitors (1–5 mM NEM, PMSF, benzamidine) and treated with bacterial collagenase (16 h, 37°C). The mixture of collagenase-resistant peptides (GRBC) was then chromatographed on two Bio-Sil TSK-250 (7.5 × 600 mm) and two Bio-Sil TSK-125 (7.5 × 600 mm) columns conected in tandem by using TSK buffer. The peptides were further purified using a Poly F reverse-phase column (DuPont) equilibrated with 0.2% trifluoracetic acid, 30% acetonitrile at 50°C and eluted with 30–80% acetonitrile gradient. SDS electrophoresis gels (Laemmli, 1979), calibrated with globular protein standards (12–95 kd, BioRad), were used to estimate the size of GRBC peptides.

Peptide sequencing and sequence comparison

The whole mixture of large GRBC peptides or individual peptides were digested with Lys-C protease (Boehringer) in 0.05 M Tris–HCl, pH 7.6, 2 M urea for 2 h at 37°C and the reaction stopped by adding acetic acid. Peptides were then separated on a Vydac C_{18} reverse-phase column equilibrated in 0.1% trifluoracetic acid at 50°C by elution with acetonitrile gradients (5–60%). Some more small peptides were obtained from a 2-fold pepsin digest of collagen VI (Chu *et al.*, 1987). Amino acid sequences were determined on a gas-phase sequencer (Applied Biosystems, model 470A) with on-line PTH amino acid analyzer (Applied Biosystems, model 120). The protein sequence databank (MIPSX Data base, F.Pfeiffer, Martinsried Institute of Protein Sequences) was searched for homologous proteins by the program FASTP (Lipman and Pearson, 1985). Alignment of two sequences was done with the program PIRALIGN (Dayhoff *et al.*, 1983). These then served as a basis for the alignment of multiple sequences.

Acknowledgements

We thank Loretta Renkart, Rui-Zhu Chang, Kerry Maddox and Bob Poljak for technical assistance. This study was supported by grants of the National Institutes of Health (grants AR-38912, AR-38923 and AR-38188), the Deutsche Forschungsgemeinschaft (SFB 1327/C4) and Shriners Hospital for Crippled children.

References

Allore,R., O'Hanlon,D., Price,R., Neilson,K., Willard,H.F., Cox,D.R., Marks,A. and Dunn,R.J. (1988) *Science*, **239**, 1311–1313.

Argraves,W.S., Deak,F., Sparks,K.J., Kiss,I. and Goetinck,P.F. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 464–468.

Aumailley,M., Mann,K., von der Mark,H. and Timpl,R. (1989) *Exp. Cell Res.*, **181**, 463–474.

Brandt,A., Glanville,R.W., Hörlein,D., Bruckner,P., Timpl,R., Fietzek,P.P. and Kühn,K. (1984) *Biochem. J.*, **219**, 625–634.

Breitbart,R.E., Andreadis,A. and Nadal-Ginard,B. (1987) *Annu. Rev. Biochem.*, **56**, 467–495.

Bruns,R.R. (1984) *J. Ultrastruct. Res.*, **89**, 136–146.

Bruns,R.R., Press,W. Engvall,E., Timpl,R. and Gross,J. (1986) *J. Cell Biol.*, **103**, 393–404.

Burgeson,R.E. (1988) *Annu. Rev. Cell Biol.*, **4**, 551–577.

Carter,W.G. (1982) *J. Biol. Chem.*, **257**, 3249–3257.

Chu,M.-L., Mann,K., Deutzmann,R., Pribula-Conway,D., Hsu-Chen,C.-C., Bernard,M.P. and Timpl,R. (1987) *Eur. J. Biochem.*, **168**, 309–317.

Chu,M.-L., Conway,D., Pan,T.-C., Baldwin,C., Mann,K., Deutzmann,R. and Timpl,R. (1988) *J. Biol. Chem.*, **263**, 18601–18606.

Colombatti,A. and Bonaldo,P. (1987) *J. Biol. Chem.*, **262**, 14461–14466.

Colombatti,A., Bonaldo,P., Ainger,K., Bressan,G.M. and Volpin,D. (1987) *J. Biol. Chem.*, **262**, 14454–14460.

Corbi,A.L., Miller,L.J., O'Connor,K., Larson,R.S. and Springer,T.A. (1987) *EMBO J.*, **6**, 4023–4028.

Cutting,G., Francomano,C.A., Chu,M.-L., Timpl,R., McCormick,M.K., Warren,A.C., Hong,H.K., Pyeritz,R.E. and Antonarakis,S.E. (1988) *Am. J. Hum. Genet.*, **43**, A141.

Dayhoff,M.D., Barker,W.C. and Hunt,L.T. (1983) *Methods Enzymol.*, **91**, 524–545.

Engel,J., Furthmayr,H., Odermatt,E., von der Mark,H., Aumailley,M., Fleischmajer,R. and Timpl,R. (1985) *Ann. N.Y. Acad. Sci.*, **460**, 25–37.

Engvall,E., Hessle,H. and Klier,G. (1986) *J. Cell Biol.*, **102**, 703–710.
Furthmayr,H., Wiedemann,H., Timpl,R., Odermatt,E. and Engel,J. (1983) *Biochem. J.*, **211**, 303–311.
Fujimura,Y., Titani,K., Holland,L.Z., Russel,S.R., Roberts,J.R., Elder,J.H., Ruggeri,Z.M. and Zimmerman,T.S. (1986) *J. Biol. Chem.*, **261**, 381–385.
Fujimura,Y., Titani,K., Holland,L.Z., Roberts,J.R. Kostel,P., Ruggeri,Z.M. and Zimmerman,T.S. (1987) *J. Biol. Chem.*, **262**, 1734–1739.
Gibson,M.A. and Cleary,E.G. (1985) *J. Biol. Chem.*, **260**, 11149–11159.
Hatamochi,A., Aumailley,M., Mauch,C., Chu,M.-L., Timpl,R. and Krieg,T. (1989) *J. Biol. Chem.*, **264**, 3494–3499.
Heckmann,M., Aumailley,M., Hatamochi,A., Chu,M.-L., Timpl,R. and Krieg,T. (1989) *Eur. J. Biochem.*, in press.
Hynes,R.O. (1987) *Cell*, **48**, 549–554.
Jander,R., Rauterberg,J. and Glanville,R.W. (1983) *Eur. J. Biochem.*, **133**, 39–46.
Jander,R., Troyer,D. and Rauterberg,J. (1984) *Biochemistry*, **23**, 3675–3681.
Keene,D.R., Engvall,E. and Glanville,R.W. (1988) *J. Cell Biol.*, **107**, 1995–2006.
Kligman,D. and Hilt,D.C. (1988) *Trends Biochem. Sci.*, **13**, 437–443.
Kozak,M. (1987) *Nucleic Acids Res.*, **15**, 8125–8147.
Kühn,K. (1984) *Collagen Rel. Res.*, **4**, 309–322.
Kuo,H.-J., Keene,D.R. and Glanville,R.W. (1989) *Biochemistry*, **28**, 3757–3762.
Laemmli,U.K. (1979) *Nature*, **227**, 680–685.
Lipman,D.J. and Pearson,W.R. (1985) *Science*, **227**, 1435–1441.
Oberbäumer,I., Laurent,M., Schwarz,U., Sakurai,Y., Yamada,Y., Vogeli,G., Voss,T., Siebold,B., Glanville,R.W. and Kühn,K. (1985) *Eur. J. Biochem.*, **147**, 217–224.
Odermatt,E., Risteli,J., van Delden,V. and Timpl,R. (1983) *Biochem. J.*, **211**, 295–302.
Pareti,F.I., Fujimura,Y., Dent,J.A., Holland,L.Z., Zimmerman,T.S. and Ruggeri,Z.M. (1986) *J. Biol. Chem.*, **261**, 15310–15315.
Pytela,R. (1988) *EMBO J.*, **7**, 1371–1378.
Rand,J.H., Patel,N., Sheng-Li,C. and Potter,B.J. (1988) *Clin. Res.*, **36**, 417A.
Rohde,H., Wachter,E., Richter,W.J., Bruckner,P., Helle,O. and Timpl,R. (1979) *Biochem. J.*, **179**, 631–642.
Ruoslahti,E. and Pierschbacher,M.D. (1987) *Science*, **238**, 491–497.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Schwarz-Magdolen,U., Oberbäumer,I. and Kühn,K. (1986) *FEBS Lett.*, **208**, 203–207.
Staunton,D.E., Marlin,S.D., Stratowa,C., Dustin,M.L. and Springer,T.A. (1988) *Cell*, **52**, 925–933.
Svoboda,K.K., Nishimura,I., Sugrue,S.P., Ninomiya,Y. and Olsen,B.R. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7496–7500.
Tikka,L., Pihlajaniemi,T., Henttu,P., Prockop,D.J. and Tryggvason,K. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7491–7495.
Timpl,R. and Engel,J. (1987) In Mayne,R. and Burgeson,R.E. (eds), *Structure and Function of Collagen Types*. Academic Press, Orlando, FL, pp. 105–143.
Titani,K. and Walsh,K.A. (1988) *Trends Biochem. Sci.*, **13**, 94–97.
Titani,K., Kumar,S., Takio,K., Ericsson,L.H., Wade,R.D., Ashida,K., Walsh,K.A., Chopek,M.W., Sadler,J.E. and Fujikawa,K. (1986) *Biochemistry*, **25**, 3171–3184.
Trüeb,B. and Winterhalter,K. (1986) *EMBO J.*, **5**, 2815–2819.
Trüeb,B., Schaeren-Wiemers,N., Schreier,T. and Winterhalter,K.H. (1989) *J. Biol. Chem.*, **264**, 136–140.
von der Mark,H., Aumailley,M., Wick,G., Fleischmajer,R. and Timpl,R. (1984) *Eur. J. Biochem.*, **142**, 493–502.
von Heijne,G. (1985) *J. Mol. Biol.*, **184**, 99–105.
von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.
Wayner,E.A. and Carter,W.G. (1987) *J. Cell Biol.*, **105**, 1873–1884.
Weber,S., Engel,J., Wiedemann,H., Glanville,R.W. and Timpl,R. (1984) *Eur. J. Biochem.*, **139**, 401–410.
Weil,D., Mattei,M.-G., Passage,E., Von Cong,N., Pribula-Conway,D., Mann,K., Deutzmann,R., Timpl,R. and Chu,M.-L. (1988) *Am. J. Hum. Genet.*, **42**, 435–445.

Received on April 3, 1989